Exploratory Data Analysis

. . . beginnings . . .

R.W. Oldford

UNIVERSITY OF WATERLOO

A confession

For a long time I have thought I was a statistician, interested in inferences from the particular to the general. But as I have watched mathematical statistics evolve, I have had cause to wonder and to doubt.

⋮

All in all I have come to feel that my central interest is in data analysis, which I take to include, among other things: procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data.

John Tukey, 1962

UNIVERSITY OF WATERLOO

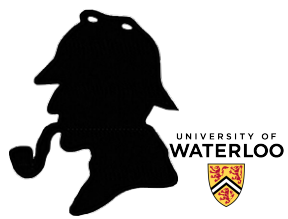
Data

"Data! data! data!" he cried impatiently. "I can't make bricks without clay."

Sherlock Holmes, from *The Adventure of the Copper Beeches* by Arthur Conan Doyle (p. 4)

Let's collect some:

1. Neatly write your name on the small **PURPLE** card
2. On one side of the card
   - Write the words "random digit"
   - Think of a "random" digit from 0 to 9
   - Write it down below the words "random digit"
3. On the other side of the same card
   - Write the words "student digit"
   - Below these, on the same side, write the **last** digit of your student id number
4. Pass the card to the end of your row, then down the rows to the front.



Data Analysis - Tukey and Wilk

"Data analysis is not a new subject. It has accompanied productive experimentation and observation for hundreds of years.

As in any other science, what is done in data analysis is very much a product of each day's technology. Every technological development of major relevance . . . has been accompanied by a tendency to rediscover the importance and to reformulate the nature of data analysis."

"The basic general intent of data analysis is simply stated: to seek through a body of data for interesting relationships and information and to exhibit the results in such a way as to make them recognizable to the data analyzer and recordable for posterity.

Its creative task is to be productively descriptive, with as much attention as possible to previous knowledge, and thus to contribute to the mysterious process called insight."

Source: J.W. Tukey and M.B. Wilk (1966) "Data Analysis and Statistics: An Expository Overview", Proc. Fall Joint Computer Conference

Data Analysis is like doing experiments

Hypothesis testing

1. Write your name on the GREEN card and turn it over.
2. Suppose we are in a licensed restaurant where patrons may drink alcohol only if they are 19 or older.
3. We have a card, one for each patron in the restaurant.
   - On one side is a picture of the beverage they are drinking.
   - On the other side is their age in years.
4. Hypothesis: No patron is illegally drinking alcohol.
   - You will be shown one side of each of four cards
   - The four cards are labelled with lower case letters (a), (b), (c), (d)
   - You may **only turn over two cards** to test the hypothesis
   - Which two do you choose to test the hypothesis?
   - You will have **only 5 seconds to choose**.
5. On the GREEN card, write down the **labels** of the two cards you selected.
6. Hand in your answers as before.

Data Analysis is like doing experiments

Hypothesis: No patron is illegally drinking alcohol.

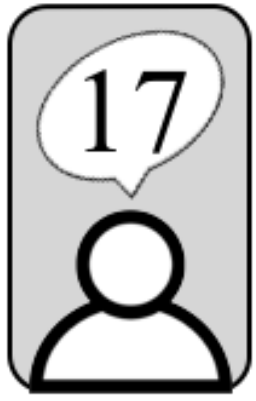
Choose two cards to turn over to test the hypothesis.

(a) (b) (c) (d)

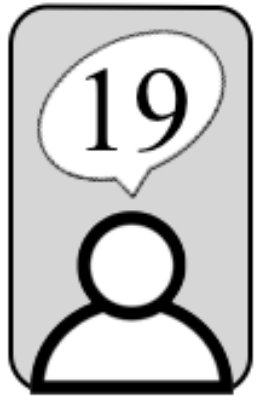
Five seconds only! Ready?

(a)

(b)



(c)



(d)

Data Analysis is like doing experiments

"The general purposes of conducting experiments and analyzing data match, point by point.

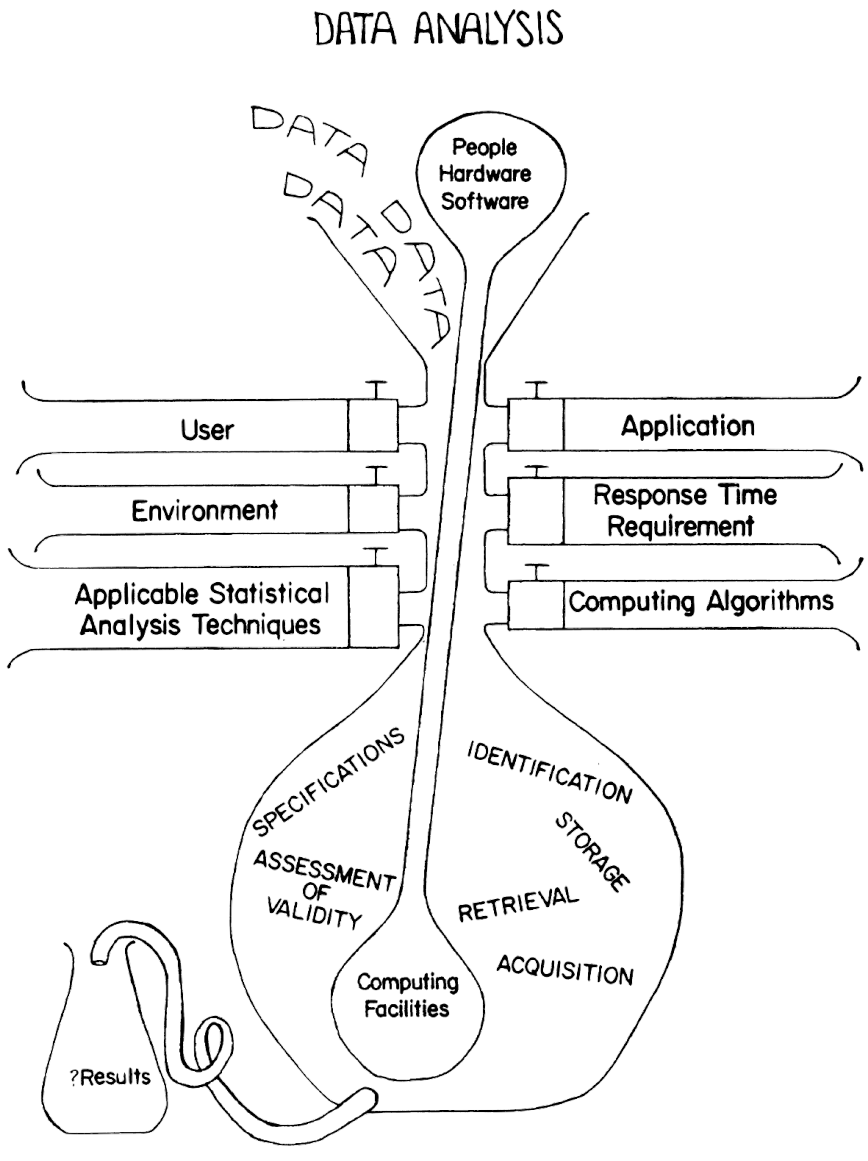
For experimentation, these purposes include

1. more adequate description of experience and quantification of some areas of knowledge;
2. discovery or invention of new phenomena and relations;
3. confirmation, or labeling for change, of previous assumptions, expectations, and hypotheses;
4. generation of ideas for further useful experiments; and
5. keeping the experimenter relatively occupied while he thinks.

Comparable objectives in data analysis are

1. to achieve more specific description of what is loosely known or suspected;
2. to find unanticipated aspects in the data, and to suggest unthought-of models for the data's summarization and exposure;
3. to employ the data to assess the (always incomplete) adequacy of a contemplated model;
4. to provide both incentives and guidance for further analysis of the data; and
5. to keep the investigator usefully stimulated while he absorbs the feeling of his data and considers what to do next.

Data Analysis is like doing experiments



Mervin Muller (1970)"Computers as an Instrument for Data Analysis", Technometrics, 12, pp. 259-293

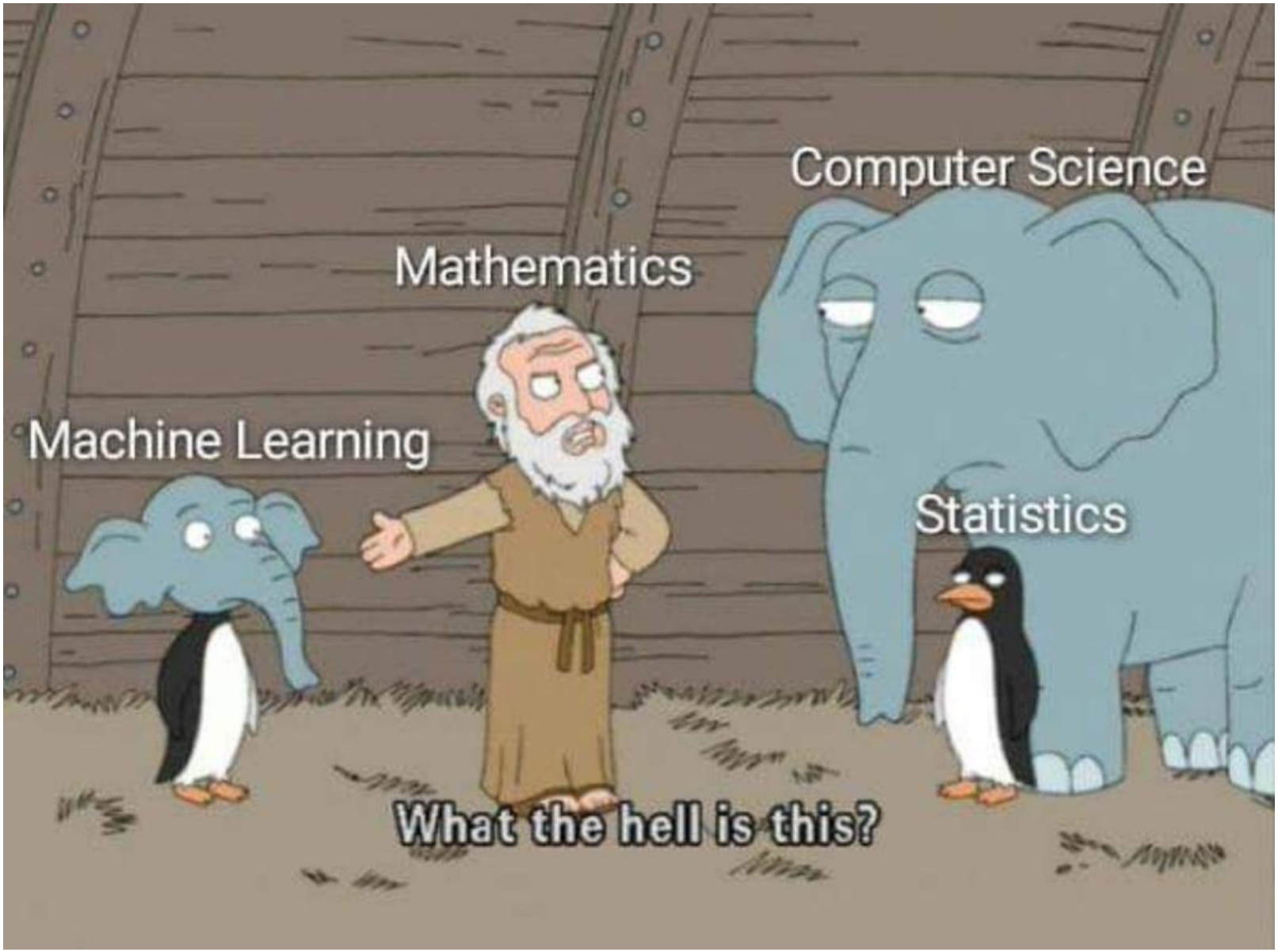
UNIVERSITY OF WATERLOO

Data Analysis – is it machine learning?

Source: xkcd: "A webcomic of romance, sarcasm, math, and language."

Data Analysis – is it machine learning?

After all . . .



Source: Family guy: "Noah's Ark."

Machine learning - the secret sauce?

The **common task framework** has these ingredients:

1. "A publicly available training dataset involving, for each observation, a list of (possibly many) feature measurements, and a class label for that observation."
2. "A set of enrolled competitors whose common task is to infer a class prediction rule from the training data."
3. "A scoring referee, to which competitors can submit their prediction rule. The referee runs the prediction rule against a testing dataset, which is sequestered behind a Chinese wall. The referee objectively and automatically reports the score (prediction accuracy) achieved by the submitted rule."

   "All the competitors share the common task of training a prediction rule which will receive a good score; hence the phase common task framework."

"It is no exaggeration to say that the combination of a predictive modeling culture together with CTF is the "secret sauce" of machine learning."

David Donoho (2017)"50 Years of Data Science", Journal of Computational and Graphical Statistics, 26, N0. 4, pp. 745-766

Examples?

Question: Is data analysis just machine learning?

UNIVERSITY OF WATERLOO

Data Analysis - characteristics shared with the "experimental process"

Among the important characteristics shared by data analysis and the experimental process are these:

1. Some prior presumed structure, some guidance, some objectives, in short some ideas of a model, are virtually essential, yet these must not be taken too seriously. Models must be used but must never be believed. As T. C. Chamberlain said, "Science is the holding of multiple working hypotheses."
2. Our approach needs to be multifaceted and open-minded. In data analysis as in experimentation, discovery is usually more exciting and sometimes much more important than confirmation.
3. It is valuable to construct techniques that are likely to reveal such complications as assumptions whose consequences are inappropriate in a specific instance, numerical inaccuracies, or difficulties of interpretation of what is found.
4. In both good data analysis and good experimentation, the findings often appear to be obvious but generally only after the fact.
5. It is often more productive to begin by obtaining and trying to explain specific findings, rather than by attempting to catalog all possible findings and explanations.

Data Analysis - characteristics shared with the "experimental process"

6. While detailed deduction of anticipated consequences is likely to be useful when two or more models are to be compared, it is often more productive to study the results before carrying out these detailed deductions.
7. There is a great need to do obvious things quickly and routinely, but with care and thoroughness.
8. Insightfulness is generally more important than so-called objectivity. Requirements for specifiable probabilities of error must not prevent repeated analysis of data, just as requirements for impossibly perfect controls are not allowed to bring experimentation to a halt.
9. Interaction, feedback, trial and error are all essential; convenience is dramatically helpful.
10. There can be great gains from adding sophistication and ingenuity – subtle concepts, complicated experimental setups, robust models, delicate electronic devices, fast or accurate algorithms – to our kit of tools, just so long as simpler and more obvious approaches are not neglected.

Data Analysis - characteristics shared with the "experimental process"

11. Finally, most of the work actually done turns out to be inconsequential, uninteresting, or of no operational value. Yet it is an essential aspect of both processes to recognize and accept this feature, with its momentary embarrassments and disappointments.

    A broad perspective on objectives and unexpected difficulties is often required to muster the necessary persistence.

In summary, data analysis, like experimentation, must be considered as an open-ended, highly interactive, iterative process, whose actual steps are selected segments of a stubbily branching, tree-like pattern of possible actions."

Data Analysis is like doing experiments

Hypothesis testing

1. Write your name on the RED card and turn it over.
2. Suppose we have a **special deck** of 52 cards
   - On one side is an upper case letter A to Z
   - On the other side is an integer between 1 and 100.
3. Hypothesis: Cards with a vowel on one side will have an even number on the other.
   - You will be shown one side of each of four cards
   - The four cards are labelled with lower case letters (a), (b), (c), (d)
   - You may **only turn over two cards** to test the hypothesis
   - Which two do you choose to test the hypothesis?
   - You will have **only 5 seconds to choose**.
4. On the RED card, write down the **labels** of the two cards you selected.
5. Hand in your answers as before.

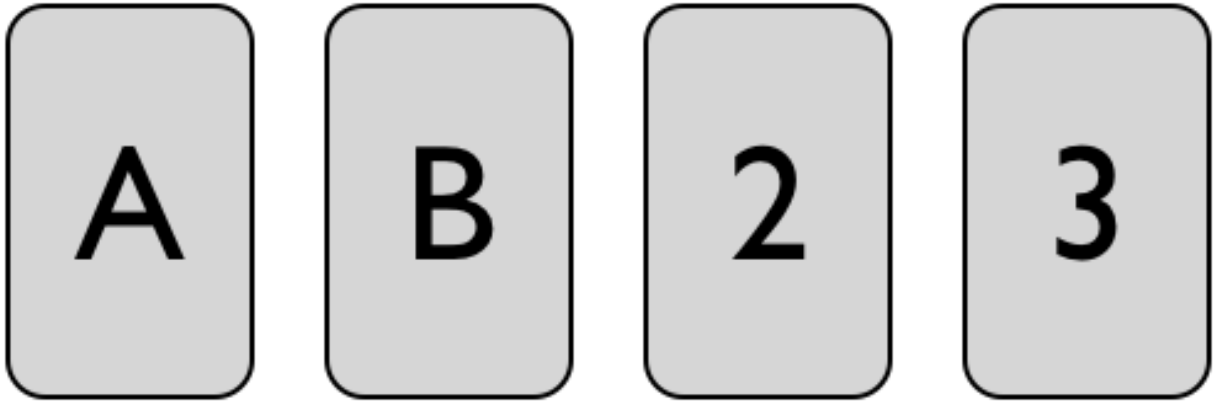
Data Analysis is like doing experiments

Hypothesis: Cards with a vowel on one side will have an even number on the other.

Choose two cards to turn over to test the hypothesis.

(a) (b) (c) (d)

Five seconds only! Ready?



(a) (b) (c) (d)

Data Analysis – a human enterprise

In relation to people:

- "The science and art of data analysis concerns the process of learning from quantitative records of experience. By its very nature it exists in relation to people. Thus, the techniques and the technology of data analysis must be harnessed to suit human requirements and talents."
- "Data analysis must be iterative to be effective. Human judgment is needed at almost every stage. (We may be able to mechanize an occasional judgment.) Unless this judgment is based on good information about what has been found, and is free to call next for what is now indicated, there cannot be really effective progress."
- "Nothing – not the careful logic of mathematics, not statistical models and theories, not the awesome arithmetic power of modern computer – nothing can substitute here for the flexibility of the informed human mind. Accordingly, both approaches and techniques need to be structured so as to facilitate human involvement and intervention."

That last paragraph really distinguishes data analysis from machine learning.

Data Analysis – summary and exposure

Key to effective data analysis: "iterative and interactive interplay of *summarizing by fit* and *exposing by residuals*".

When "sufficiently arithmetic" the relationship can be expressed additively as

$$observation = fit + residual$$

or perhaps as multiplicatively

$$observation = fit \times (residual \;\; factor).$$

E.g. in most "supervised learning", *fit* is the *prediction* and *residual* its *error*.

Think more generally of *fit* as *summary* and *residual* as *exposure* of the *unexplained leftovers*.

Tukey and Wilk:

"[Fitting] is helpful in summarizing, exposing and communicating. Each fit gives a summary description, provides a basis for exposure based on residuals, and may have the parsimony needed for effective communication."

"Adequate examination of residuals is one of the truly incisive tools of exposure."

Data Analysis – diverse objectives of fitting

Again, think of fit as summary, prediction, description, model, explanation, . . .
Objectives of the fit:

1. *Pure description.* e.g. a simple numerical or graphical summary
2. *Local prediction.* Substitute new values of the input x into model to get predicted value of y.
   - "hoping that our description of the past, however empirical, will continue to be a good description of the future."
3. *Global prediction of local change.* i.e. assess how moderate changes in x affect the prediction of y.
   - "If this is to be accomplished successfully, the general situation must be favorable, and theory, or insight, or broad experience must have been responsible for choosing the form of the fit and the nature of the y variables; the data before us can rarely be used to narrow things down enough to provide such good prediction, even of changes, elsewhere."
4. *Global prediction of values.* i.e. predicting y given an x far outside the range of the data.
   - "Reliance upon outside information (including insight) is now even greater, and the chances of success are correspondingly diminished."
5. *Tell which variables have influence.* "This is sometimes possible, but nowhere nearly as often as is commonly hoped. Very frequently several alternative sets of variables will each give a satisfactory fit."
6. *Estimate coefficients as "physical constants".* This is much rarer.

Data Analysis – diverse objectives of fitting

Again, think of residual as the leftover, remains after fitting, the unexplained, . . .
Objectives of the residual:

1. *Immediate use for further summarization*, providing
   - values for further study – e.g. a time series *after* it has been seasonally adjusted
   - a basis for immediate further fitting – e.g. adding a quadratic term to a straight line fit
2. *Emphasize exposure*, "examining and exposing with a view to
   - learning about the *inadequacy of the fit*"
   - identifying *peculiar values*" (and structure) "either for study in their own right or for suppression, partial or complete, from further analysis."

Suppression? "Data is often dirty. Unless the dirt is either removed or decolorized, it can hide much that we would like to learn."

Suppression

- "may be complete and wholly human, as when we decide to exclude a particular set of observations from all computations."
- "may be partial and wholly automatic, as when we use the median of a set of observations"
- "In practice, the processes of selection and suppression are mixed up, rather complex, and for all that quite essential."

Data Analysis – strategy

"In addition to the two-pronged use of summarization and exposure, including careful attention to residuals, three of the main strategies of data analysis are:

1. Graphical presentation.
   - For large-scale data analysis, there is really no alternative to plotting techniques, properly exploited. A picture is not merely worth a thousand words, it is much more likely to be scrutinized than words are to be read.
2. Provision of flexibility in viewpoint and in facilities.
   - We must have flexibility in the choice of a model for summarization, in the selection of the data to be employed, in computing the summary, in choosing the fitting procedures to be used, and in selecting the terms in which the variables are to be expressed.
   - Flexibility in assembly and reassembly of techniques is crucial.
3. Intensive search for parsimony and simplicity, including careful reformation of variables and bending the data to fit simple techniques.
   - The importance of parsimony in data analysis can hardly be overstated. By parsimony we mean both the use of few numerical constants and also the avoidance of undue complexity of form in summarizing and displaying. The need for parsimony is both aesthetic and practical.

As a matter of general strategy we may note here that it is almost always easier, and usually better, to "unbend" data to fit known analysis techniques than to bend the techniques to fit the data. . . . With enough effort we can probably bend any of our techniques of data analysis to work explicitly and effectively on the data in its raw form, but this effort is rarely justified."

Data Analysis – general characteristics

"In productive data analysis:

1. *Those who seek are more likely to find.*
   - Tight frameworks of probable inference demand advance specifications of both a model and a list of questions to be asked, followed by data collection and then by analysis intended to answer only the prechosen questions. A man who lived strictly by this paradigm would have a hard time in learning anything new.
2. *Flexibility in viewpoint and in facilities is essential for good data analysis.*
   - Data analysis is very much a bootstrap exercise. Our facilities and our attitudes must encourage flexibility: use of alternate models, choice of subsets of the data, choice of subsets of auxiliary or associated variables, choice of forms of expression of these variables, and of the data, choice of alternate criteria, both in fitting and in evaluating.
3. *Both exploration and description are major objectives of data analysis; for both reasons data analysis is intrinsically iterative.*
   - Both the search for insight and for the unanticipated require that the available information be displayed. *Description as a preparation to display and insight is, in a certain sense, the main business of data analysis.* But, equally, adequate insight, however informal or intuitive, is a necessary precursor for incisive description of the anticipated. Accordingly, insightful exploration and description require an iterative, interactive, complementary process involving both summarization and exposure."

Data Analysis – guidance and models

"In the sense in which we here use the word "model" – a means of guidance without implication of belief or reality – *all the structures that guide data analysis*, however weak and nonspecific, are models – even when they are not explicitly mathematical.

"Definiteness in detailing objectives and assumptions in a formal model can simplify mathematical problems and increase the simplicity and impact of the results reached. But tightness of detail usually forces such a formal model unnecessarily far away from the realities of the data-gathering situation, obscuring possibly important phenomena. Looser structures can often do as well in simplicity and clarity of results while retaining robustness and breadth. *Both for guidance and the encouragement of exploration, it is most desirable that models be loose and noncommittal*, thus encouraging diverse alternative working hypotheses.

"Trying to answer questions concerning the adequacy of a model by the use of data may be interesting and valuable. *In data analysis, however, models and techniques are to be thought of and developed as assisting tools with the focus on the data.* The models need not fit perfectly or even adequately to prove usefully insightful. We must never believe so deeply in any model as to constrain our insight."

Data Analysis – limitations

"Data analysis cannot make knowledge grow out of nothing or out of mere numbers, nor can it salvage or sanctify poor work. *It can only bring to our attention a combination of the content of the data with the knowledge and insight about its background which we must supply. Accordingly, validity and objectivity in data analysis is a dangerous myth.*"

"Developing models of one sort to aid in appreciating or assessing the performance of models of another sort (perhaps describing methods of analysis) may indeed be useful to the discipline of data analysis. Such theories of inference must, however, be taken only as a guidance, and kept from becoming impediments. Assumptions and theory are indispensable, but, in use, the focus of data-analysis techniques must be on the data and the analysis, with the theory aiding insight by providing alternative backdrops.

"It seems too easy for some to believe that detailed assumptions can make the data tell much more, either qualitatively or quantitatively, than would otherwise be the case. But if these assumptions are unwarranted their consequences may be misleading.

Data Analysis – a science

"There are diverse views as to what makes a science, but three constituents will be judged essential by most, viz:

a. intellectual content,
b. organization in an understandable form,
c. reliance upon the test of experience as the ultimate standard of validity.

By these tests mathematics is not a science, since its ultimate standard of validity is an agreed-upon sort of logical consistency and provability.

As I see it, data analysis passes all three tests, and I would regard it as a science, one defined by a ubiquitous problem rather than by a concrete subject. Data analysis and the parts of statistics which adhere to it, must then take on the characteristics of a science rather than those of mathematics,

⋮

These points are meant to be taken seriously."

John Tukey, 1962

Data Analysis – technology

"The basic purpose of a technology is to provide and organize tools and techniques to meet relatively well-specified, but often very broad, objectives.

"By the term data analysis we mean to encompass the techniques, attitudes, interests and concepts which are relevant to the process of learning from organized records of experience. This area has always been of fundamental importance.

"In thinking about data analysis technology, the antithesis between hardware and software (between machinery and organized know-how) is important, not only in the conventional uses of these terms within a computer system, but also in data analysis itself. Specific techniques, such as a three-way analysis of variance, considered as parts of data analysis, are really data-analytic hardware. The mystery and art of when and why to use such techniques make up the data-analytic software, which is today very soft indeed.

"Our difficulties are twice compounded; we must develop data-analytic software to harness the power of our data-analytic hardware and, at the same time, develop computer software for data analysis that adequately harnesses the power of our computer and display hardware."

Tukey and Wilk, 1966.

Data Analysis Environments - their evolution

"Viewed in retrospect, statistical computing shows an overall trend, beginning from small scale, ad hoc and inflexible programs to large-scale efforts aiming at a broad range of analysis, and exhibiting considerable complexity and flexibility.

"This evolution is reinforced, rather than obscured, by the parallel changes in computers and their software accompaniment and in the requirements and sophistication of statistical techniques.

"With improved hardware, larger-scale problems in data analysis can be tackled.

"The programmer is relieved of some of the most tedious considerations of efficiency. . . . The contribution of better programming languages and monitors is even more important.

"The development and distribution of _____ in the past decade is surely a prime factor in the revolution in scientific data processing.

"Not only is the actual computing power greatly increased, but the ability to communicate programs and techniques widely is introduced."

Exercise: Give some detailed examples of these comments; e.g. fill in the blank.

Data Analysis Environments - their evolution

A coarse classification of data analysis environments as they evolved

1. One-off isolated special purpose programs
   - write a program in some language to solve the problem at hand
   - oldest, simplest, and in many cases, the most common approach
   - Advantage: get results **now** for **this** problem
   - Disadvantages:
     - excessive programming labour due to duplication of effort
     - inconsistency across programs with respect to data conventions, etc.
     - inflexibility of resulting code, algorithms, etc. making reuse difficult
2. Collections of pre-packaged routines
   - common important tasks are coded up as individual routines
   - removes some of the programming effort by using routines as building blocks
     - e.g. subroutine libraries
   - Advantages:
     - less tedious and error-prone programming
     - keep flexibility of "doing one's own programming"
   - Disadvantages:
     - **requires** further programming by the analyst and **understanding** of the routines
     - "depends on the intelligence of the designers and their ability to isolate the essential structure in the data analysis"
   - "There is a fundamental trade-off between good detailed numerical algorithms and computational building blocks which fit together easily to produce finished programs"
   - "The search for a near optimum in this regard is one of our most serious

3. Collection of pre-determined fixed analyses

- sort of an "electronic statistician"
- large scale systems providing 'one-stop shopping' for a large number of data analysis problems
- Advantages:
  - empowers those who do not have the time, staff, or background to take advantage of previous approaches
  - often wholly subsumes the pieces of the two previous classifications
- Disdvantages:
  - forces the analyst into pattern matching the problem to the available solutions
  - not easily adapted to unorthodox solutions which need to combine or modify the existing analyses
  - often requires a "very rigid form of command, expressed in an arbitrary and error-prone coding system"
  - choosing the correct encoding can amount to programming, thus defeating the purpose of this approach

4. Towards a statistical computing language

- cannot anticipate the demands that will be made on the system
- provide a set of tools that can be put together to construct statistical procedures *without outside programming*
- must be able "to reuse the results of of analyses, to make decisions between analyses in flight and to add new links to the systems."
- *Advantages:*
  - *if the "language" is truly expressive for statistical reasoning about data, then the data analyst will indeed have a powerful tool at their disposal.*
- *Challenges:*
  - *capturing the basic statistical concepts (or "nouns"), and statistical actions/procedures (or "verbs") well can be extremely challenging (e.g. how should data be represented? models? methods?)*
  - *integrating "language" with visual and direct manipulation (e.g. mouse gestures)*
  - *does it scale?*

Italicized comments are mine. See also my 1987 abstract statistical computing for some extension and further thoughts along this direction.

Data Analysis Environments - their evolution

"There has come a point in the design of some recent systems when the concept of predetermined fixed analyses begins to give way to that of blocks or links combined by the user to produce an analysis for his problem.

The standard procedures are still there, but they need not be used in a standard way.

The actual boundary between this and the previous classification is arbitrary but the difference is crucial.

In a sense, the designer has reassumed the more humble position of the first two approaches.

He specifically assumes that he may not be able to anticipate the demands on his system, and provides instead useful tools to construct statistical procedures.

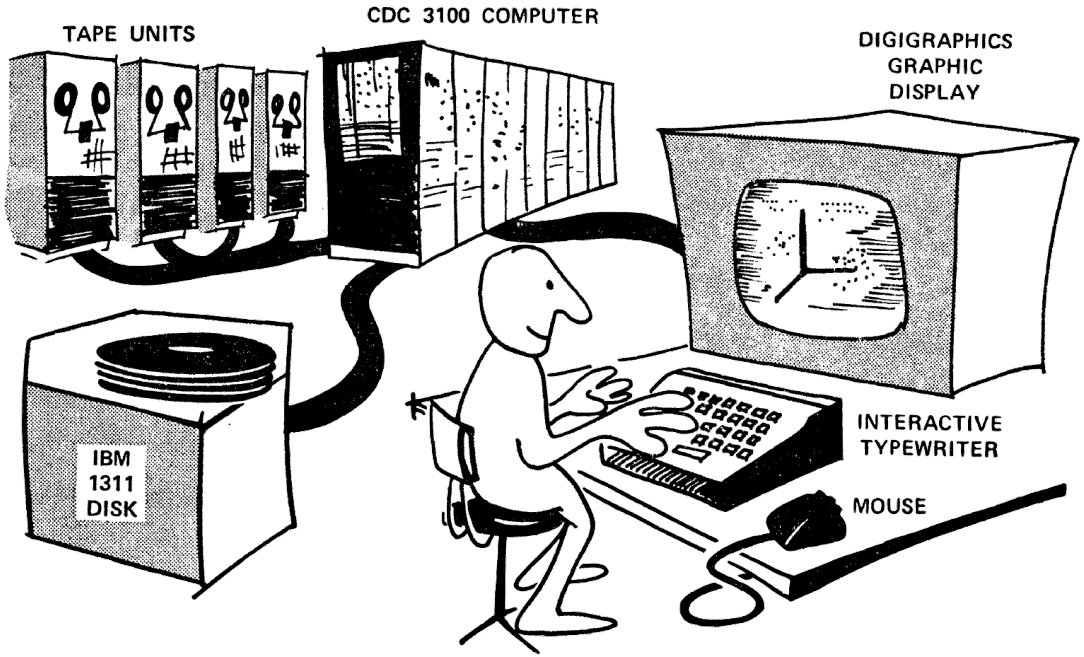
The contrast with the second classification is that the tools can be put together to make up the standard analyses without outside programming."

Exercise: Give some detailed examples of these comments.

Source: John M. Chambers (1967) "Some General Aspects of Statistical Computing", Journal of the Royal Statistical Society. Series C (Applied Statistics), Vol. 16, No. 2(1967), pp. 124-132

Data Analysis Environment - interactive

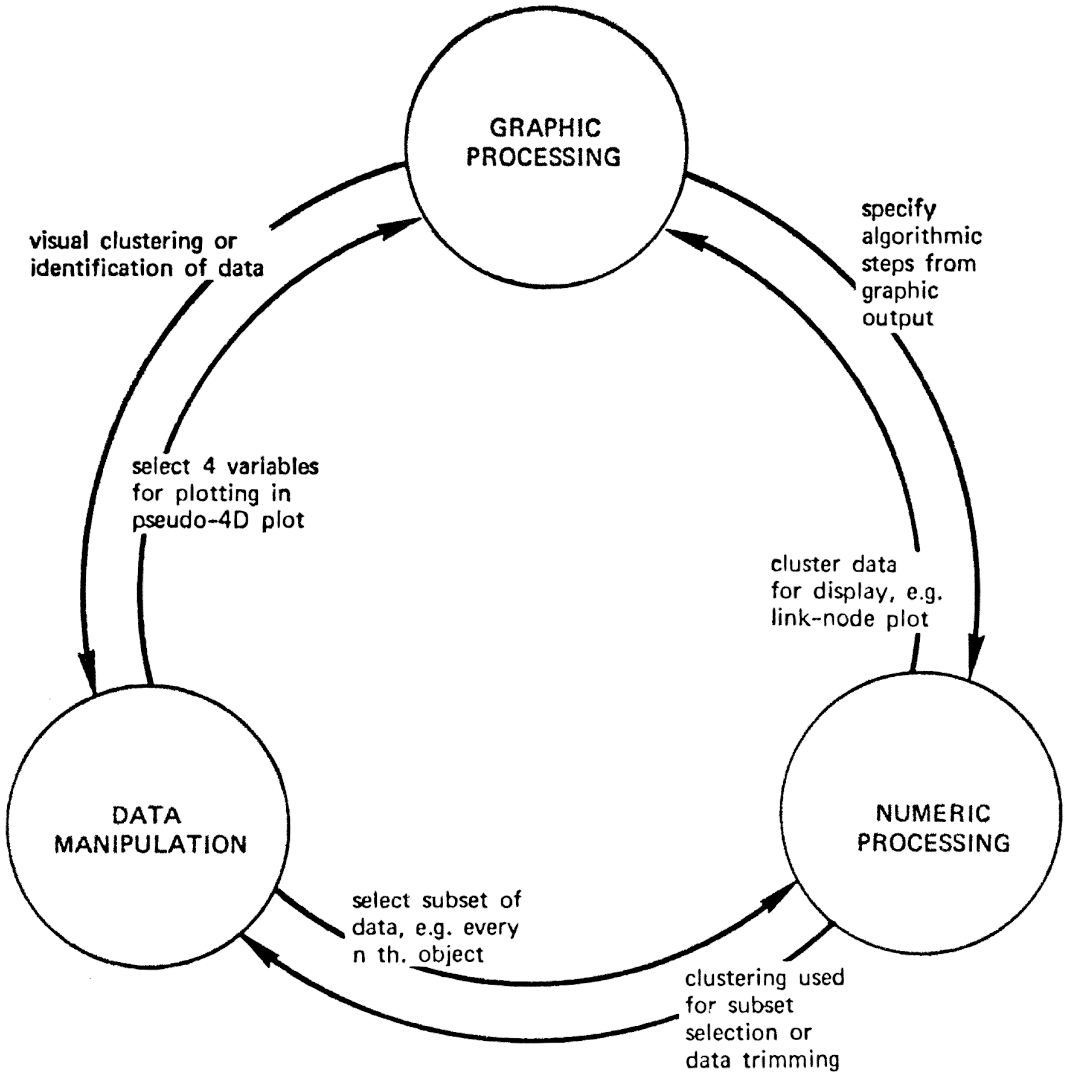
One of the earliest interactive data analysis environments (PROMENADE):



Geoffrey H. Ball and David J. Hall (1970) "Some Implications of Interactive Graphic Computer Systems for Data Analysis and Statistics", Technometrics, Vol. 12, No. 1 (Feb., 1970), pp. 17-31

Data Analysis Environment - interactive

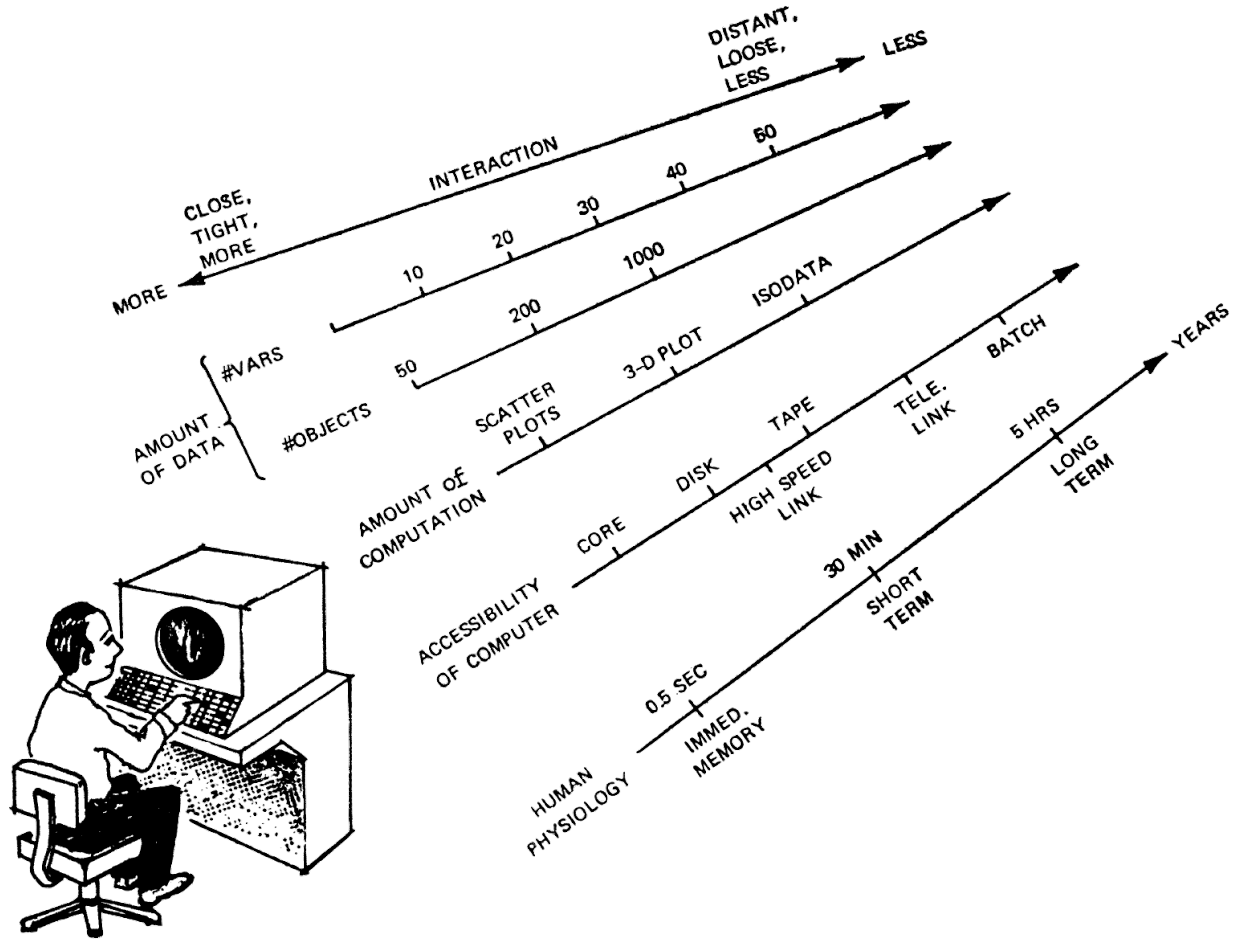
Principal activities of their (PROMENADE) system:



Geoffrey H. Ball and David J. Hall (1970) "Some Implications of Interactive Graphic Computer Systems for Data Analysis and Statistics", Technometrics, Vol. 12, No. 1 (Feb., 1970), pp. 17-31

Data Analysis Environment - interactive

Some of the challenges:

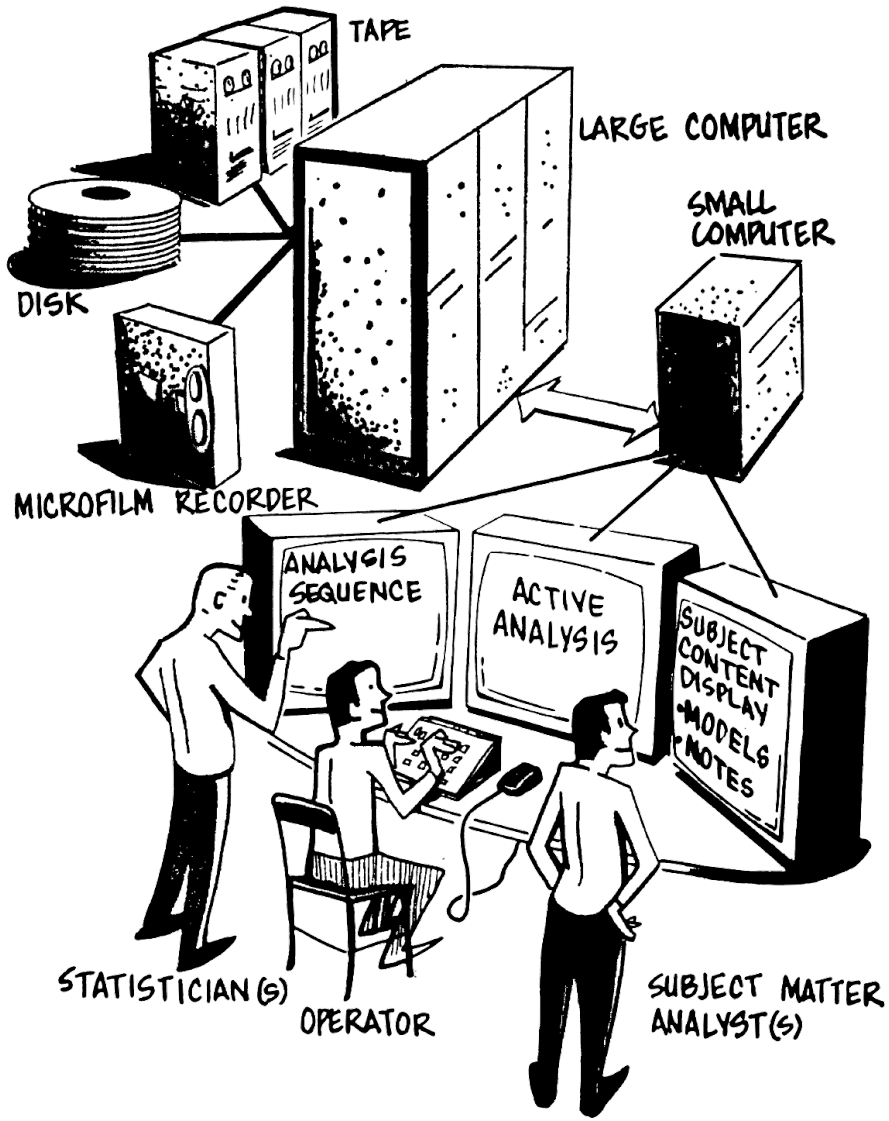


Geoffrey H. Ball and David J. Hall (1970) "Some Implications of Interactive Graphic Computer Systems for Data Analysis and Statistics", Technometrics, Vol. 12, No. 1 (Feb., 1970), pp. 17-31

UNIVERSITY OF WATERLOO

Data Analysis Environment - interactive

Future interactive analysis system:



Geoffrey H. Ball and David J. Hall (1970) "Some Implications of Interactive Graphic Computer Systems for Data Analysis and Statistics", Technometrics, Vol. 12, No. 1 (Feb., 1970), pp. 17-31

Data Analysis Environment - S language (from 1976 . . . 1990s)

J.M. Chambers, R.A. Becker, Allan Wilks, and many others (see Rick Becker's A Brief History of S which covers 1976 to the 1990s) developed the statistical programming language at Bell Laboratories called

Several books and many papers were written about S over its many incarnations. E.g. see the Communications of the ACM paper on the S language.

Data Analysis Environment

As should be clear, there is a **very long** history of statistical data analysis and of data analysis environments.

Lots of research over 50 years on interactive data analysis systems.

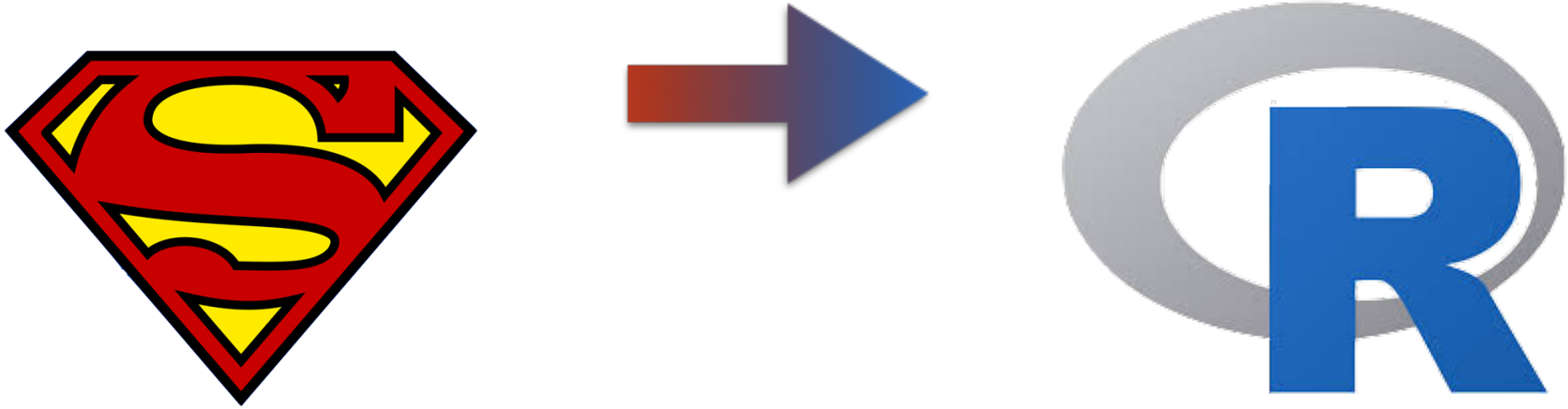
- graphics workstations for example
  - do it yourself . . . PRIM-9, PRIM-H, PRIM-A, . . .
  - SUNs, Symbolics Lisp machines, Xerox lisp machines, . . .
  - Macs/PCs: MacSpin, DataDesk, Battelle research
  - BLITs (Bell Labs Intelligent Terminals)
- many interactive systems
  - S, ISP, BLSS, DataDesk, DINDE (video), LispStat, SysStat, Quail, Mondrian (video), . . .
  - graphics models: "canvas", "plot windows", "data viewer" and its descendants XGobi and ggobi, "views", . . .
- statistical strategy (e.g. implementation and study, software abstraction) and intelligent data analysis
  - intelligent statistical software and statistical expert systems (e.g. Artificial Intelligence, AI and Statistics, COMPSTAT 86, 88), REX, rule-based statistical expert systems, STUDENT, DINDE, Quail (see also Computational thinking), . . .
- reproducible research
  - history mechanisms, Data Analysis Networks, Statistical Analysis Maps, history graphs, graphical programming, . . .

UNIVERSITY OF WATERLOO

See also the ASA Video Library for some historical examples.

Data Analysis Environment - R: S language goes open source

In the 90s S begins being reimplemented as R and around 1997 the R language goes open source



Many researchers contribute to the project (see R contributors).

R has become the de facto descendant of all of this research and experience (e.g. unlike python).

Book series:

- Springer's Use R! series
- R books
- R wikibook
- Google's Rstyle guide

Journals:

- R Journal

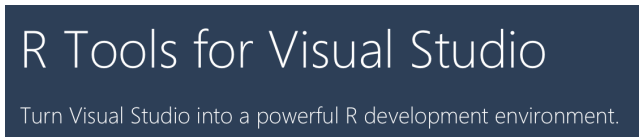
Conferences:

- R Conferences

Data Analysis Environment - R programming environments

Lots of companies, and lots of integrated development environments (IDEs)

Visual Studio

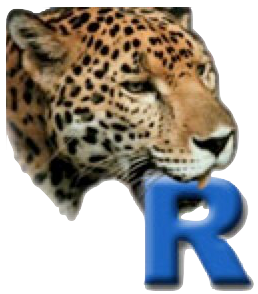


Emacs Speaks Statistics

Eclipse and R

Microsoft/Revolution

JGR



R from Jupyter

We will be using RStudio

Some relevant `R` packages:

- call `C` or `Fortran` is part of `base`
- `Rcpp` interface to `C++` code
- call `Java` via `rJava`
- call `tcl/tk` via `tcltk` package
- call `python` via `rPython` package
- packages to support parallel programming, e.g. `parallel`
- can use `Hadoop` from `R`
- can call `spark` from `R` via the `sparklyr` package
- can access various data bases (e.g. `Oracle`, `mySQL`, `postgreSQL`, `DB2`)
- reproducible research via `Rmarkdown` (e.g. `knitr` and other packages) or `RNotebook`

PLUS lots of more interfaces worth exploring.